联合结构化图学习与 l_1 范数谱嵌入的鲁棒聚类算法 *

汤立伟，张家珲，彭　勇†，孔万增
(杭州电子科技大学 计算机学院，杭州 310018)

摘　要： 谱聚类算法一般是在给定的输入图上进行谱分解，然后通过后置处理(如 K 均值聚类或谱旋转)得到最终的聚类结果。此类方法存在两个不足：a)将图的构造与谱分解割裂成两个独立的阶段，导致了结果的次优性；b)常用的基于 l_2 范数度量谱特征向量的相似性具有噪声敏感性。为了克服上述两点不足，提出基于联合结构化图学习与 l_1 范数谱嵌入的鲁棒聚类算法(记为 CLRL1)。在该算法框架下，一方面图的学习过程与聚类过程可以有效结合起来进行协同优化，另一方面 l_1 范数的使用可以很好地约束谱特征向量的相似性以提升算法的鲁棒性。在多个常用数据集上进行的实验结果表明，改进的算法聚类性能得到了明显的提升。

关键词： 谱聚类；结构化图学习；l_1 范数；联合学习
中图分类号： TP18　　doi: 10.19734/j.issn.1001-3695.2020.01.0034

Robust clustering algorithm based on joint structured graph learning and l_1-norm spectral embedding

Tang Liwei, Zhang Jiahui, Peng Yong†, Kong Wanzeng
(*School of Computer Science & Technology*, *Hangzhou Dianzi University*, *Hangzhou* 310018, *China*)

Abstract: Given a fixed graph, the graph-based clustering usually performs the eigen-decomposition to obtain the spectral eigenvectors based on which we need to conduct post-processing steps such as Kmeans or spectral rotation to obtain the final clustering assignments. This paradigm may cause two limitations: a) this two-stage strategy breaks down the connection between the graph construction and the calculation of spectral eigenvectors; and b) the l_2-norm based similarity measure between spectral eigenvectors is usually sensitive to noise. To deal with these two limitations, this paper proposed a robust clustering algorithm based on joint structured graph learning and l_1-norm spectral clustering, termed CLRL1. In the proposed framework, on one hand the graph learning process and the clustering process can be optimized together towards the optimum; on the other hand, the l_1-norm similarity measure of spectral eigenvectors is used to improve the model robustness. Experiments on extensive benchmark data sets show the effectiveness of the proposed algorithm.

Key words: spectral clustering; structured graph learning; l_1-norm; joint learning

0　引言

聚类是机器学习与数据挖掘领域一类重要的数据分析方法并得到了广泛的应用，例如图像分割[1]和复杂网络分析。聚类是一个将数据样本划分为多个类簇的过程，其目标是使得簇内样本的相似度较高，而簇间样本的相似度较低。常用的聚类方法有k均值聚类[2, 3]，基于图的谱聚类方法[4~7]等；一般情况下，两类方法都可以取得较好的聚类效果。

传统的谱聚类方法都需要经过两个阶段，第一阶段是基于样本数据关系图的构造，第二阶段是采用优化手段对图拉普拉斯矩阵进行谱分解以得到聚类标志矩阵。通常这样获得的聚类标志矩阵是连续的并且可能含有负值，需要再通过k均值方法或者是谱旋转方法来进行离散化已得到最终的聚类结果。这种两阶段的方式将图的构造与谱分解割裂成两个独立的过程，造成了最终结果的次优性。因此，传统谱聚类算法具有以下两个缺陷：a) 无法从图中直接获得最终的聚类结果，还需采用其他的聚类方法以获得聚类结果。b) 谱聚类结果依赖于构造图的质量，对特定图的构造方法极其敏感[8]。

ℓ_2范数是传统谱聚类算法普遍运用于度量数据间相似度的度量指标，它具有噪声敏感性，一般情况下仅适用于对高斯噪声进行建模。相比之下，ℓ_1范数[9]更适合于稀疏的大噪声，可以提高模型的鲁棒性[10]。

本文提出了一种新型基于结构化图学习的ℓ_1范数谱嵌入聚类方法，能有效弥补了传统谱聚类的缺陷。本文在真实数据集和加噪数据集上做了大量实验，结果表明本文提出的新型聚类算法的表现超过同类算法，具有良好的收敛性和鲁棒性。

1　算法的提出

图聚类学习一般是在给定的图上进行，假定与图相对应的相似矩阵为$\mathbf{W} \in \mathbb{R}^{n\times n}$[11](这里$n$为数据点的个数)。如果基于数据样本构造的图$\mathbf{W}$质量不高，将直接导致聚类的效果变得不理想。本文的目标是在图$\mathbf{W}$的基础上学习一个新的具有良好特性的结构化图$\mathbf{S} \in \mathbb{R}^{n\times n}$。显然，$\mathbf{S}$应该是非负的[12]，并且行和为 1 的。除此之外，本文基于如下的定理来给出关于$\mathbf{S}$的秩的约束[13, 14]。

定理 1　拉普拉斯矩阵中零特征值的个数等于图关联矩阵对应的图中连通分量的个数。

本文希望$\mathbf{S}$具有秩k(假定聚类的类簇个数为k)，根据上述定理，将其转换为对应的拉普拉斯矩阵的秩的约束为$rank(\mathbf{L_S}) = n - k$。这里$\mathbf{L_S} = \mathbf{D_S} - (\mathbf{S} + \mathbf{S})/2$，其中$\mathbf{D_S}$为对角的度矩阵，第$i$个对角元素定义为$(\mathbf{D_S})_{ii} = \sum_i s_{ij}$。基于图关联矩阵$\mathbf{W}$，进行结构化图$\mathbf{S}$的学习，可以通过优化如下的

收稿日期： 2020-01-15；**修回日期：** 2020-04-16　**基金项目：** 国家自然科学基金资助项目(61971173，61602140)；浙江省科技计划资助项目(2017C33049)；中国博士后科学基金资助项目(2017M620470)；浙江省新苗人才计划资助项目(2019R407030)

作者简介： 汤立伟(1999-)，男，浙江苍南人，本科生，主要研究方向为机器学习和模式识别；张家珲(1998-)，男，浙江绍兴人，本科生，主要研究方向为机器学习与模式识别；彭勇(1985-)，男(通信作者)，安徽巢湖人，副教授，硕导，博士，主要研究方向为机器学习、模式识别与脑机交互系统.(yongpeng@hdu.edu.cn)；孔万增，男，教授，博导，博士，主要研究方向为脑机交互.

目标函数来实现，其中$\mathbf{S1}$表示矩阵$\mathbf{S}$与向量$\mathbf{1}$相乘，$\mathbf{S1}=\mathbf{1}$即约束矩阵$\mathbf{S}$行和为 1。

$$\min_{\mathbf{S}\geq \mathbf{0},\mathbf{S1}=\mathbf{1}} \|\mathbf{S}-\mathbf{W}\|_2^2,\ s.t.\ rank(\mathbf{L_S})=n-k. \tag{1}$$

假定$\sigma_i(\mathbf{L_S})$是$\mathbf{L_S}$的第i小特征值，鉴于图拉普拉斯矩阵具有半正定性质，显然$\sigma_i(\mathbf{L_S})\geq \mathbf{0}$；因此(1)式可等价于

$$\min_{\mathbf{S}\geq \mathbf{0},\mathbf{S1}=\mathbf{1}} \|\mathbf{S}-\mathbf{W}\|_2^2+\gamma\sum_{i=1}^{k}\sigma_i(\mathbf{L_S}). \tag{2}$$

当正则化参数γ足够大时，(1)式中的约束条件近似被(2)式中的第二项满足。假设数据$\mathbf{x}_i\in R^{d\times 1}$对应的聚类标志向量为$\mathbf{f}_i\in R^{1\times k}$，根据 Ky Fan 定理[14]，(2)式可以写成

$$\min_{\mathbf{S},\mathbf{F}}\|\mathbf{S}-\mathbf{W}\|_2^2+\gamma\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}\|\mathbf{f}^i-\mathbf{f}^j\|_2^2, \quad s.t.\ \mathbf{S}\geq \mathbf{0},\mathbf{S1}=\mathbf{1},\mathbf{F}\in\mathbb{R}^{n\times k},\mathbf{F}^T\mathbf{F}=\mathbf{I}_k. \tag{3}$$

这里的$\mathbf{F}$为聚类标志矩阵，其第i行对应于$\mathbf{f}^i$。

可以看到，式(3)中的第二项，对于聚类标志向量$\mathbf{f}^i$和$\mathbf{f}^j$相似性的度量使用的是ℓ_2范数。一般来说，ℓ_2范数比较适合于刻画高斯类误差，而且是噪声敏感的。这里本文采取文献[9]中提出的ℓ_1范数[15]，来提高模型的鲁棒性[15, 16]，即采用如下的基于ℓ_1范数的谱嵌入模型：

$$\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}}\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}\|\mathbf{f}^i-\mathbf{f}^j\|_2. \tag{4}$$

构造一个n^2维度的向量$\mathbf{P}$，使得$\mathbf{P}$的第$((i-1)*(n+j))$个元素为$s_{ij}\|\mathbf{f}^i-\mathbf{f}^j\|_2$。最小化$\mathbf{P}$的$\ell_1$范数会使最终得到的目标变得更加稀疏，从而产生一个更利于聚类结果的$\mathbf{F}$。所以式(4)实际上是一个广义的ℓ_1范数，这里本文沿用文献[9]的命名。

结合文献[8, 9]的思想，本文提出最终的优化算法模型：

$$\min_{\mathbf{S},\mathbf{F}}\|\mathbf{S}-\mathbf{W}\|_2^2+\gamma\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}\|\mathbf{f}^i-\mathbf{f}^j\|_2, \quad s.t.\ \mathbf{S}\geq \mathbf{0},\mathbf{S1}=\mathbf{1},\mathbf{F}^T\mathbf{F}=\mathbf{I}_k. \tag{5}$$

该目标函数的正则化项是非平滑的，本文提出了一种迭代的算法去求解，即交替优化图关联矩阵$\mathbf{S}$与聚类标志矩阵$\mathbf{F}$，直至目标函数收敛为止。如图 1 所示，传统的谱聚类先构图后进行谱分解，本文提出的 CLRL1 算法联合进行图的构造与谱分解，并且采用了结构化图学习与ℓ_1范数普嵌入的方法以提高模型的鲁棒性。

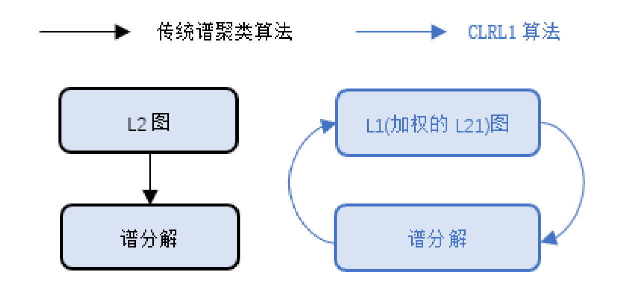


图 1　CLRL1 算法与传统谱聚类算法的流程对比

Fig. 1　Comparison of CLRL1 and traditional spectral clustering algorithm

2　模型的优化

为优化目标函数式(5)，本文在固定一个变量的情况下求解另一个变量。以下是每个变量更新规则的详细推导过程。

a) 固定$\mathbf{F}$更新$\mathbf{S}$。问题式(5)中与$\mathbf{S}$关联的目标函数为

$$\min_{\mathbf{S}}\sum_{i=1}^{n}\sum_{j=1}^{n}(s_{ij}-w_{ij})^2+\gamma\sum_{i,j=1}^{n}d_{ij}s_{ij}, \quad s.t.\ s_{ij}\geq 0,\sum_{j=1}^{n}s_{ij}=1. \tag{6}$$

这里$d_{ij}=\|\mathbf{f}^i-\mathbf{f}^j\|_2$，即$\mathbf{d}_i$是一个向量，其第$j$个元素为$d_{ij}$(同理得$\mathbf{s}_i$和$\mathbf{w}_i$)。对$\mathbf{s}_i$进行配方，可得：

$$\min_{\mathbf{s}_i}\frac{1}{2}\left\|\mathbf{s}_i-(\mathbf{w}_i-\frac{\gamma}{2}\mathbf{d}_i)\right\|_2^2, s.t.\ \mathbf{s}_i\geq \mathbf{0},\mathbf{s}_i^T\mathbf{1}=1. \tag{7}$$

该式实际上对应一个单纯形约束下的欧式距离问题。令$\mathbf{v}_i=\mathbf{w}_i-\frac{\gamma}{2}\mathbf{d}_i$，(7)式对应的拉格朗日函数为

$$\mathcal{L}(\mathbf{s}_i)=\frac{1}{2}\|\mathbf{s}_i-\mathbf{v}_i\|_2^2-\eta(\mathbf{s}_i^T\mathbf{1}-1)-\mathbf{s}_i^T\beta. \tag{8}$$

其中$\eta\in\mathbb{R}$和$\boldsymbol{\beta}\in\mathbb{R}^{n\times 1}$为拉格朗日乘子。本文将$\mathbf{s}_i$具体的求解算法总结在算法 1 中，具体的推导过程可参见文献[17]。

算法 1　固定$\mathbf{F}$求得$\mathbf{S}$的算法

输入：给定的向量$v_i\in R^{n\times 1}$；

输出：目标向量$s_i\in R^{n\times 1}$。

a)计算$g=v_i-\frac{11^T}{c}v_i+\frac{1}{c}1$；

b)根据牛顿法得到根λ^*；

c)对于每个$t\in[1,c]$，得到最优的$s_i^{(t)*}=(g_t-\lambda^*)_+$。

b) 固定$\mathbf{S}$更新$\mathbf{F}$。关于变量$\mathbf{F}$的拉格朗日函数为

$$L(F)=\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}\|f^i-f^j\|_2-Tr(\Lambda(F^TF-I)). \tag{9}$$

定义$\widetilde{\mathbf{S}}$是$\mathbf{S}$的重加权的图关联矩阵，即

$$\tilde{s}_{ij}=\frac{s_{ij}}{2\|\mathbf{f}^i-\mathbf{f}^j\|_2}, \tag{10}$$

其对应的拉普拉斯矩阵为$\widetilde{\mathbf{L}}=\widetilde{\mathbf{D}}-\widetilde{\mathbf{S}}$，$\widetilde{\mathbf{D}}$是一个对角矩阵，第$i$行的对角元素是$\sum_{j=1}^{n}\tilde{s}_{ij}$。使用式(9)对$\mathbf{F}$求导并令导数为零，可以得到

$$\frac{\partial\mathcal{L}(\mathbf{F})}{\partial\mathbf{F}}=\widetilde{\mathbf{L}}\mathbf{F}-\mathbf{F}\boldsymbol{\Lambda}=\mathbf{0}. \tag{11}$$

显然，聚类标志矩阵$\mathbf{F}$的解为矩阵$\widetilde{\mathbf{L}}$前 k 个小特征值对应的特征向量所组成的矩阵。因为$\widetilde{\mathbf{L}}$是依赖于$\mathbf{F}$，因此需要迭代的更新$\mathbf{F}$和$\widetilde{\mathbf{L}}$。

本文将目标函数(5)的整个优化过程总结在算法 2 中。

算法 2　联合结构化图学习与ℓ_1范数谱嵌入的鲁棒聚类算法

输入：数据$\mathbf{X}\in\mathbb{R}^{d\times n}$，类簇个数$k$，正则化参数$\gamma$。

输出：聚类标志矩阵$\mathbf{F}\in\mathbb{R}^{n\times k}$。

a) 初始化。根据 HeatKernel 函数来计算图关联矩阵$\mathbf{W}\in\mathbb{R}^{n\times n}$(邻域为 5，带宽参数设置为数据点对距离的平均值)；计算对应的拉普拉斯矩阵$\mathbf{L}$,并对其进行特征分解以初始化$\mathbf{F}$。

b)当算法未收敛时执行：固定$\mathbf{F}$，根据算法 1 更新$\mathbf{S}$；根据式(10)计算$\widetilde{\mathbf{S}}$与$\widetilde{\mathbf{L}}$；固定$\mathbf{S}$，根据式(11)更新$\mathbf{F}$；

下面对算法优化过程的收敛性进行简要分析。CLRL1 模型的求解分为两部分，一部分是求解$\mathbf{S}$，另一部分是求解$\mathbf{F}$。由于$\mathbf{S}$是解析解，故只需证明$\mathbf{F}$的收敛性[18]。

引理 1　对于任意非零向量$\mathbf{f}_t\in\mathbb{R}^k$，有如下不等式成立

$$\|\mathbf{f}\|_2-\frac{\|\mathbf{f}\|_2^2}{2\|\mathbf{f}_t\|_2}\leq\|\mathbf{f}_t\|_2-\frac{\|\mathbf{f}_t\|_2^2}{2\|\mathbf{f}_t\|_2}. \tag{12}$$

证明　显然$(\sqrt{f}-\sqrt{f_t})^2\geq 0$，故有

$$(\sqrt{f}-\sqrt{f_t})^2\geq 0 \implies f-2\sqrt{ff_t}+f_t\geq 0 \implies \sqrt{f}-\frac{f}{2\sqrt{f_t}}\leq\frac{\sqrt{f_t}}{2} \implies \sqrt{f}-\frac{f}{2\sqrt{f_t}}\leq\sqrt{f_t}-\frac{f_t}{2\sqrt{f_t}} \tag{13}$$

将式(13)中的f和f_t使用$\|\mathbf{f}\|_2^2$和$\|\mathbf{f}_t\|_2^2$进行代换，即得到式(12)。

根据引理1，本文给出如下关于算法1收敛性的定理。

定理2 CLRL1算法将在每次迭代中单调减小目标函数式(5)第二项的值，并收敛到问题的局部最优值。

证明 当$\mathbf{S}$固定时，根据CLRL1算法的执行步骤5，可以得到

$$\mathbf{F}_{t+1}=\arg\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{s}_{ij}\|\mathbf{f}^i-\mathbf{f}^j\|_2^2. \tag{14}$$

其中$\tilde{s}_{ij}=\frac{s_{ij}}{2\|\mathbf{f}^i-\mathbf{f}^j\|_2}$，所以

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{s_{ij}\|\mathbf{f}_{t+1}^i-\mathbf{f}_{t+1}^j\|_2^2}{2\|\mathbf{f}_t^i-\mathbf{f}_t^j\|_2}\le\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{s_{ij}\|\mathbf{f}_t^i-\mathbf{f}_t^j\|_2^2}{2\|\mathbf{f}_t^i-\mathbf{f}_t^j\|_2}. \tag{15}$$

根据引理1，可以得到

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}(\|\mathbf{f}_{t+1}^i-\mathbf{f}_{t+1}^j\|_2-\frac{\|\mathbf{f}_{t+1}^i-\mathbf{f}_{t+1}^j\|_2^2}{2\|\mathbf{f}_t^i-\mathbf{f}_t^j\|_2})\\&\le\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}(\|\mathbf{f}_t^i-\mathbf{f}_t^j\|_2-\frac{\|\mathbf{f}_t^i-\mathbf{f}_t^j\|_2^2}{2\|\mathbf{f}_t^i-\mathbf{f}_t^j\|_2})\end{aligned} \tag{16}$$

将不等式(15)和(16)两边相加，可得

$$\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}\|\mathbf{f}_{t+1}^i-\mathbf{f}_{t+1}^j\|_2\le\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}\|\mathbf{f}_t^i-\mathbf{f}_t^j\|_2. \tag{17}$$

根据不等式(17)，在t轮迭代过后，算法将收敛。达到收敛后，式(17)左右两边会达到相等，此时，$\mathbf{F}_t$和$\tilde{\mathbf{L}}_t$将满足式(11)，即满足KKT条件。因此算法CLRL1的收敛性即得到了证明。

3 实验结果

为检验所提出算法的有效性，本文分别在非加噪和加噪的数据集上进行了聚类。下面分别从数据与实验设置、实验结果与分析等方面进行介绍。为了评估聚类的结果，本文利用精度(ACC)，标准化互信息(NMI)和纯度(Purity)三个指标来衡量各个算法的表现[19]。

3.1 数据集与实验设置

为了比较算法的性能，选择在标准数据集上进行测试。使用的是COLI20灰度物体图片数据集，Yeast酵母数据集，Wine葡萄酒化学成分数据集，Uspst手写数字数据集，AR人脸数据集，Dig手写数字数据集。

COLI20数据集包含了1440张128×128像素的灰度图片(包含了 20个不同物体，每个物体每隔5度进行一次拍摄)，分为了20类。

Yeast数据集包含了1484条预测蛋白质定位的10类数据样本。

Wine数据集包含了178条酒的化学分析数据，每条数据样本有13个维度，共3类。

Uspst数据集是Usps数据集的子集，每张手写数字图是16×16像素的大小，一共2007张图片，一共是10类手写数字。

AR数据集包含2600张60×43像素的100类彩色人脸图片，图像具有正面视图面，具有不同的面部表情，照明条件和遮挡(太阳眼镜和围巾)，每个人参与两次拍摄，相隔两周(14天)，两次都拍摄相同的照片。AR为来源于100个人的人脸图像数据集，每个人分两次共拍摄了26张人脸图像，单次拍摄13张(其中7张为不同光照、不同表情等条件，3张为戴眼镜，3张为戴围巾)。在前一部分实验中，对于每个人的人脸图像，本文选择不含眼镜与围巾装饰的14张，由此形成一个1400张的数据子集。对于第二部分加噪实验中，本文使用所有的2600张人脸图像作为数据。

Dig数据集包含了1797张8×8像素的0-9手写数字灰度图片。各数据集的样本数、维度与类簇数特性如表1所示。

表1　所选数据集的描述

Tab. 1　Description of the selected datasets

数据集	样本数	维度	类簇数
COLI20	1440	1024	20
Yeast	1484	1470	10
Wine	178	13	3
Uspst	2007	256	10
AR	2600	2580	100
Dig	1797	64	10

选择比较的算法是K-means, NMF, NCut, CLR[8], L1_un[9], RMNMF[19]。CLR是一种图学习的聚类算法，L1_un是一种利用L1范数图的聚类算法。RMNMF是一种联合非负矩阵分解与谱聚类的算法。对于聚类模型CLRL1、CLR、L1_un、NCut、NMF、RMNMF本文设置近邻数为5，权重使用热核函数(HeatKernel function)来进行计算，其中带宽参数使用各点对之间距离的平均值。对于各NMF算法，基矩阵的列数设为聚类类别数。如果模型中存在自由正则化参数，按照$\{10^{-5},10^{-4}\cdots,10^{5}\}$的次序挑选出使聚类结果最优的参数进行保留。根据文献[19]，RMNMF对正则化参数λ在0.001～1之间的选择不敏感，本文将RMNMF算法中的自由参数设置为0.01。对于本文提出的CLRL1算法以及CLR算法，本文采用了一种启发式的方法来加速调参过程，先将正则化参数λ设置为一个小的值，在之后的每一次迭代中，如果计算得到$\mathbf{L_S}$的0特征值个数大于k，就将λ除以2，如果计算得到$\mathbf{L_S}$的0特征值个数小于k，就将λ乘2，否则就停止迭代，得到最佳λ。对于k均值聚类和需要k均值聚类进行后处理的NMF算法、Ncut算法，本文将k均值聚类算法重复5次，记录下得到的最佳结果。

3.2 非加噪数据实验结果

表2~4分别给出了参与比较的各个算法在选用的6个数据集上的所得到的结果，其中最好的结果在表格中以加粗形式表示。

表2　各算法的聚类精度对比

Tab. 2　Performance of different algorithms with respect to accuracy /%

	COLI20	Yeast	Wine	Uspst	AR	Dig
Kmeans	45.28	13.68	70.22	63.03	16.15	70.78
NMF	49.93	23.79	58.43	62.73	35.95	64.50
Ncut	68.13	24.26	61.80	59.34	14.36	77.35
CLR	78.47	29.72	72.47	54.46	34.59	66.33
L1_un	72.08	30.19	72.47	68.31	26.52	70.56
RMNMF	59.51	36.32	**73.59**	68.06	33.95	75.79
CLRL1	**85.21**	**37.53**	72.47	**70.65**	**37.17**	**81.74**

表3　各算法的归一化互信息聚类指标对比

Tab. 3　Performance of different algortihms with respect to normalized mutual information /%

	COLI20	Yeast	Wine	Uspst	AR	Dig
Kmeans	72.54	1.98	**42.88**	60.50	50.15	69.66
NMF	70.47	6.56	33.37	60.04	65.80	66.37
Ncut	79.63	6.42	34.68	64.11	43.46	84.36
CLR	93.53	4.43	39.48	70.87	58.92	75.71
L1_un	90.18	3.53	39.48	78.73	55.77	85.10
RMNMF	68.39	**13.44**	39.61	60.19	61.86	67.04
CLRL1	**95.77**	13.36	39.48	**81.94**	**75.45**	**89.39**

表 4　各算法的聚类纯度表现

Tab. 4　Performance of different algorithms with respect to Purity /%

	COLI20	Yeast	Wine	Uspst	AR	Dig
Kmeans	49.10	32.48	70.22	70.80	17.16	73.79
NMF	55.00	33.76	62.92	68.61	38.81	70.90
Ncut	70.14	34.37	61.80	63.73	19.30	81.30
CLR	85.00	32.68	72.47	64.97	37.74	67.84
L1_un	72.30	32.35	72.47	71.90	27.16	70.56
RMNMF	60.48	**41.91**	**73.59**	72.84	35.74	75.79
CLRL1	**89.79**	39.42	72.47	**80.92**	**67.12**	**81.75**

可以看出，本文改进的 CLRL1 算法在所用的数据集上取得了较好的结果，聚类效果明显优于其他的算法。对精度指标，CLRL1 算法在 COLI20、Yeast、Uspst、AR、Dig 数据集上取得了最优的结果，分别比次优算法精度提高 6.74%，1.21%，2.34%，2.58%和 4.39%。RMNMF 与 CLRL1 是两种不同的聚类算法，RMNMF 使用 L21 范数度量矩阵分解的误差，是一种联合矩阵分解与谱聚类的算法，可以看成是特征学习与聚类同时进行；CLRL1 重点在于联合结构化图学习与谱聚类，二者侧重点有所不同，因而对数据以及噪声分布特性的表现有所不同，虽然 CLRL1 在少数数据集的某些指标上未取到最优结果，但在这些情况下二者实验结果的差别比较微弱，多数情况下 CLRL1 的表现更优。由此本文认为将图的构造过程与聚类标志矩阵的求解过程联合起来进行协同优化以避免两阶段模式带来的次优性对聚类性能提升是有益的。虽然此实验中未对数据集进行主动加噪，但是数据集本身也存在一定的噪声，因此改进的 CLRL1 算法相对于 CLR 性能也有一定的改善。

图 2 给出了 CLRL1 模型中的正则化参数在 Uspst 数据集上的敏感性实验结果。对该数据集，γ的取值设置在[10,50]将是合适的。对其他数据集，参数的敏感性曲线有类似的走势。

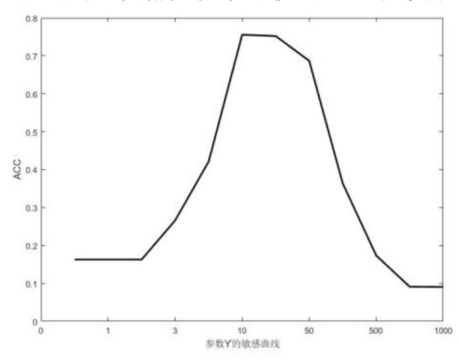


图 2　在 Uspst 数据集上参数γ的敏感曲

Fig. 2　Sensitive curve of parameter γ on the Uspst dataset

关于 CLRL1 算法的收敛性，图 3 给出了其在 COLI20 和 Uspst 两个数据集上目标函数值随迭代次数增加的下降情况。可以看出，该算法具有很好的收敛速度。

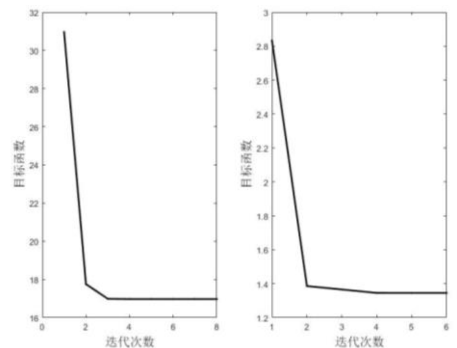


图 3　在 COLI20 和 Uspst 数据集上目标函数的收敛曲线图

Fig. 3　Convergence property of objective function value on COLI20 dataset and Uspst dataset

3.3　加噪数据实验结果

为了测试的 CLRL1 的鲁棒性，本节的实验将对 COLI20 数据集做主动加噪处理，即在原始图像中及嵌入块状噪声和椒盐噪声，如图 4 所示。块状加噪后分别得到在原图像上随机加入 1、2 块 4×4 随机位置噪声块的数据集，椒盐加噪分别得到在原图像上加上 10%、20%随机噪点的数据集，如图 4 所示。此外，本文还将在包含围巾与眼镜遮挡的 AR 人脸图像数据集上进行实验；图 5 给出了单个人员分两次拍摄的 26 张人脸图片。

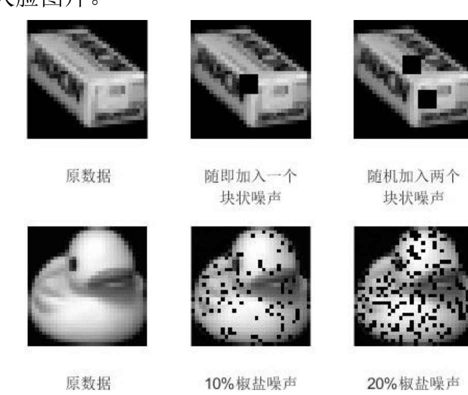


图 4　COLI20 数据集非加噪图和加噪样图

Fig. 4　Non-noisy and noisy sample images in COLI20 data set



图 5　单个人两次实验拍摄的 26 张人脸图像

Fig. 5　Sample face images of one subject in two sessions of AR data set

本文将对本文提出的算法 CLRL1 与 CLR 算法、L1_un 算法、RMNMF 算法分别在四个人工加噪数据集与加噪 AR 数据集上进行实验。图 6~8 给出了三种算法在含块状躁声的 COLIL20 数据集上的实验结果；图 9~11 给出了三种算法在随机像素置乱 COIL20 数据集上的实验结果。表 5 给出了对比算法在含眼镜与围巾遮挡的完整 AR 数据集上的实验结果。

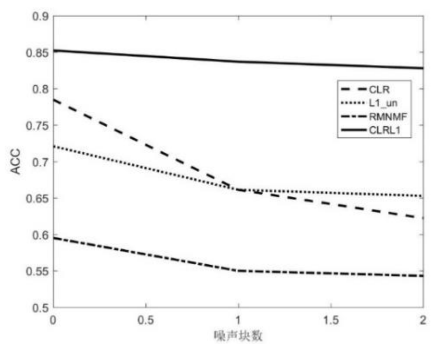


图 6　在含块状噪声 COIL20 数据集上的各算法聚类精度

Fig. 6　ACC of algorithms on COIL20 with block occlusion

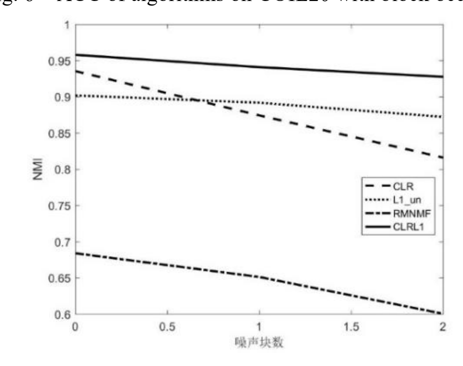


图 7 在含块状噪声 COIL20 数据集上的各算法聚类互信息

Fig. 7　NMI of algorithms on COIL20 with block occlusion

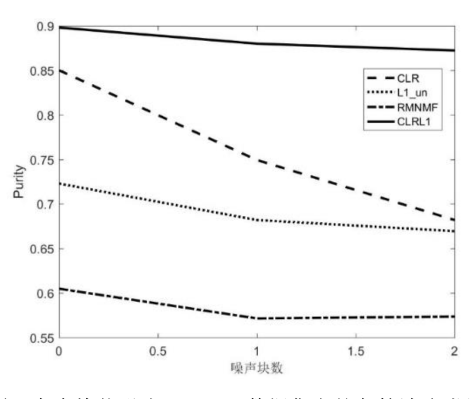


图 8 在含块状噪声 COIL20 数据集上的各算法聚类纯度

Fig. 8　Purity of algorithms on COIL20 with block occlusion

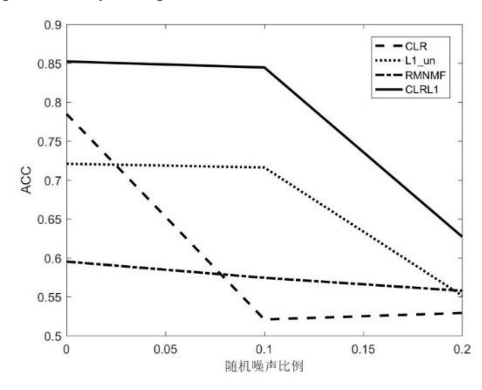


图 9　在随机像素置乱 COIL20 数据集上的各算法聚类精度

Fig. 9　ACC of algorithms on COIL20 with random pixel corruption

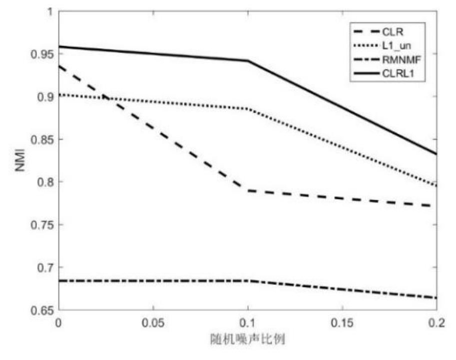


图 10 在随机像素置乱 COIL20 数据集上的各算法聚类互信息

Fig. 10　NMI of algorithms on COIL20 with random pixel corruption

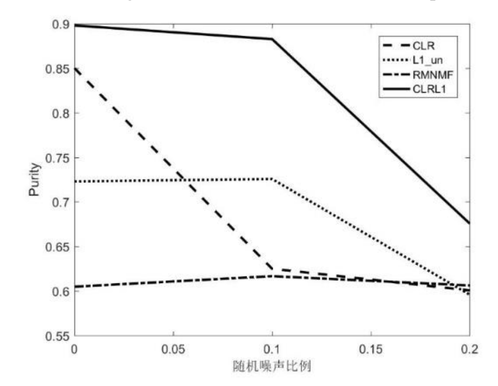


图 11 在随机像素置乱 COIL20 数据集上的各算法聚类纯度

Fig. 11　Purity of algorithms on COIL20 with random pixel corruption

表 5　在含真实噪声的 AR 数据集上各算法的聚类效果

Tab. 5　Performance of algorithms on AR data set with real disguises

	ACC	NMI	Purity
CLR	0.1854	0.3802	0.2000
L1_un	0.1535	0.3292	0.1600
RMNMF	0.1992	0.4698	0.2107
CLRL1	**0.2015**	**0.5126**	**0.3777**

从上述实验结果，可以看出本文提出的 CLRL1 算法鲁棒性上明显优于传统 CLR 算法、L1_un 算法以及 RMNMF 算法。本文使用的 L1 范数是广义的 L1 范数，即由加权的 L21 范数相加得到的。这使得 L1 范数使得算法在面对噪声时的表现更加鲁棒[19, 20]。对比 CLR 与 CLRL1 的实验结果，本文认为，在随机像素置乱、块状噪声与真实噪声的数据集上，使用$\ell 1$范数来度量聚类标识向量之间的相似性均可以提升模型的鲁棒性。对比 L1_un 与 CLRL1 的实验结果，本文可以得出联合实现图的学习与聚类过程要优于两阶段模式的“先构图再聚类”的效果。

4　结束语

本文提出了一种联合结构化图学习与$\ell 1$范数谱嵌入的鲁棒聚类算法。一方面，该算法将图的学习与聚类过程结合起来，在同一个目标函数中实现二者的协同优化，避免了“先构图再聚类”这一两阶段模式带来的次优性问题；另一方面，该算法使用了$\ell 1$范数来度量聚类标识向量之间的距离，避免了常用的$\ell 2$范数度量带来的噪声敏感问题。通过大量的实验(正常数据集与加噪数据集)上的实验结果表明，改进的算法具有更好的鲁棒性，聚类效果得到了提升。

参考文献：

[1] 刘汉强，张青．区间模糊谱聚类图像分割方法 [J]．计算机工程与科学, 2018, 40 (09): 1611-1616. (Liu Hanqiang, Zhang Qing. Interval fuzzy spectral clustering image segmentation method [J]. Computer Engineering and Science, 2018, 40 (09): 1611-1616.)

[2] Wagstaff K, Cardie C, Rogers S, *et al.* Constrained k-means clustering with background knowledge [C]// Proc of the Eighteenth International Conference on Machine Learning. New York: ACM Press, 2001, 1: 577-584.

[3] 刘靖明，韩丽川，侯立文．基于粒子群的 K 均值聚类算法 [J]．系统工程理论与实践, 2005, 25 (6): 54-58. (Liu Jingming, Han Lichuan, Hou Liwen. K-means clustering algorithm based on particle swarm [J]. Systems Engineering-Theory & Practice, 2005, 25 (6): 54-58.)

[4] 蔡晓妍，戴冠中，杨黎斌．谱聚类算法综述 [J]．计算机科学, 2008, 35 (7): 14-18. (Cai Xiaoyan, Dai Guanzhong, Yang Libin. Overview of spectral clustering algorithms [J]. Computer Science, 2008, 35 (7): 14-18.)

[5] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm [C]// Advances in Neural Information Processing Systems. 2002: 849-856.

[6] Chan P K, Schlag M D F, Zien J Y. Spectral k-way ratio-cut partitioning and clustering [J]. IEEE Trans on Computer-Aided Design of Integrated Circuits and Systems, 1994, 13 (9): 1088-1096.

[7] 万月，陈秀宏，何佳佳．基于加权密度的自适应谱聚类算法 [J]．计算机工程与科学, 2018, 40 (10): 1897-1901. (Wan Yue, Chen Xiuhong, He Jiajia. Adaptive spectral clustering algorithm based on weighted density [J]. Computer Engineering and Science, 2018, 40 (10): 1897-1901.)

[8] Nie Feiping, Wang Xiaoqian, Jordan M I, *et al.* The constrained laplacian rank algorithm for graph-based clustering [C]// Thirtieth AAAI Conference on Artificial Intelligence. 2016: 1969-1976.

[9] Nie Feiping, Wang Hua, Huang Heng, *et al.* Unsupervised and semi-supervise learning via $\ell 1$-norm graph [C]// IEEE International Conference on Computer Vision. 2011: 2268-2273.

[10] 邹鹏，李凡长．鲁棒谱多流形聚类算法 [J]．小型微型计算机系统, 2018, v. 39 (06): 12-16. (Zou Peng, Li Fanchang. Robust spectral multi-manifold clustering algorithm [J]. Mini-Micro Systems, 2018, v. 39 (06):

12-16.)

[11] 王卫卫, 李小平, 冯象初, 等. 稀疏子空间聚类综述 [J]. 自动化学报, 2015, 41 (8): 1373-1384. (Wang Weiwei, Li Xiaoping, Feng Xiangchu, *et al.* Overview of Sparse Subspace Clustering [J]. Acta Automatica Sinica, 2015, 41 (8): 1373-1384.)

[12] Li Tao, Ding C. The relationships among various nonnegative matrix factorization methods for clustering [C]// IEEE International Conference on Data Mining. 2006: 362-371.

[13] Chung F R K, Graham F C. Spectral graph theory [M]. American Mathematical Society, 1997.

[14] Fan K. On a theorem of Weyl concerning eigenvalues of linear transformations I [J]. Proceedings of the National Academy of Sciences of the United States of America, 1950, 36 (1): 31-35.

[15] Ding C, Zhou Ding, He Xiaofeng, *et al.* R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization [C]// International Conference on Machine Learning. 2006: 281-288.

[16] Peng Yong, Lu Baoliang. Robust structured sparse representation via half-quadratic optimization for face recognition [J]. Multimedia Tools and Applications, 2017, 76 (6): 8859-8880.

[17] Nie Feiping, Yang Sheng, Zhang Rui, *et al.* A general framework for auto-weighted feature selection via global redundancy minimization [J]. IEEE Transactions on Image Processing, 2019, 28 (5) , 2428-2438.

[18] Nie Feiping, Huang Heng, Cai Xiao, *et al.* Efficient and robust feature selection via joint ℓ2, 1-norms minimization [C]// Advances in neural information processing systems. 2010: 1813-1821.

[19] Huang Jing, Nie Feiping, Huang Heng, *et al.* Robust manifold nonnegative matrix factorization [J]. ACM Transactions on Knowledge Discovery from Data, 2014, 8 (3): 11: 1-21.

[20] Zhao Haifeng, Wang Zheng, Nie Feiping. A new formulation of linear discriminant analysis for robust dimensionality reduction [J]. IEEE Transactions on Knowledge and data engineering, 2018, 31 (4): 629-640.